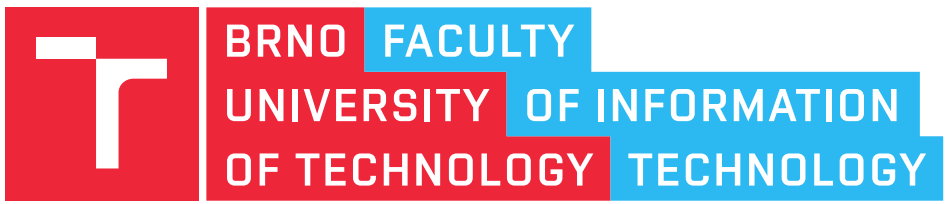
Modèles de sous-espaces pour la découverte d'unités acoustiques

Lucas Ondel, Bolaji Yusuf

Université technique de Brno, Faculté des technologies de l'information
Božetěchova 1/2. 612 66 Brno - Královo Pole
{iondel,xyusuf00}@fit.vutbr.cz



December 10, 2020

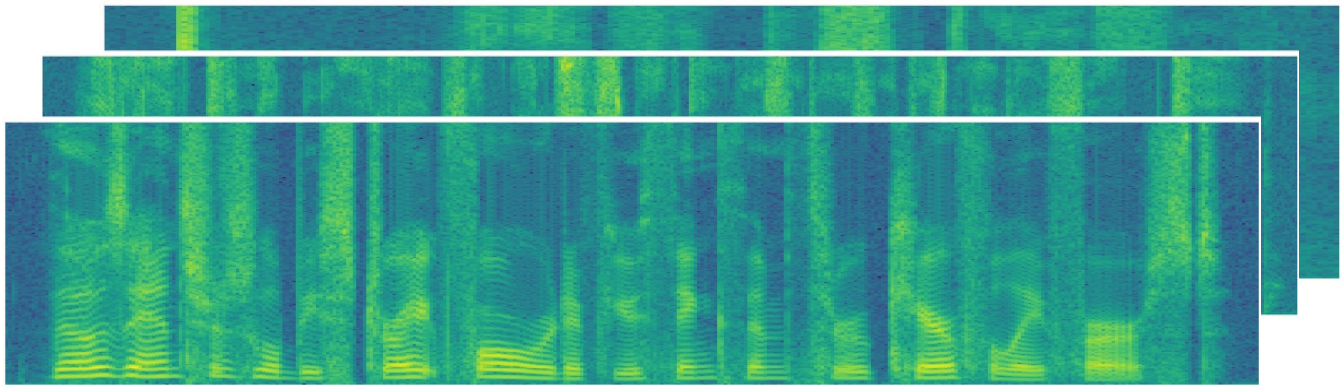
Table des matières

Table des matières

Table des matières

Définition

- En entrée: enregistrements audios sans transcription
- En sortie:
  - un inventaire de simili-phones (appelés “unités acoustiques”)
  - segmentation et transcriptions



Apprendre comme un bébé

Table des matières

- La diversité culturelle diminue à l'échelle mondiale
  - technologies de la parole sont disponible pour seulement une poignée de langue
  - une majorité des langues sont orales: impossible d'appliquer les technologies actuelles
- Un système efficace de DUA pourrait:
  - aider les linguistes à documenter langues à risques
  - servir de bloc de base pour construire des technologies de la parole pour les langues non écrites
- 2022-2032: Décennie des langues autochtones ▸ UNESCO website

- Les processus cognitifs d'apprentissage chez l'humain sont mal connus:
  - le cerveau est un organe complexe
  - l'apprentissage de la parole se passe alors que l'enfant ne peux pas communiquer verbalement
- Approche par rétro-ingénierie ("E. Dupoux. 2018") Link:
  - construisons un système capable d'apprendre la parole de manière non supervisée
  - analyse du système pour en tirer des principes généraux sur l'apprentissage

- Le "toujours plus de données" soulève de nombreux problèmes :
  - sociaux/éthiques: monopole des technologies de l'apprentissage par les propriétaires de **large** base données
  - écologiques: plus de données demande plus d'énergie
- La recherche sur la DUA implique nécessairement des modèles d'apprentissage économes en données

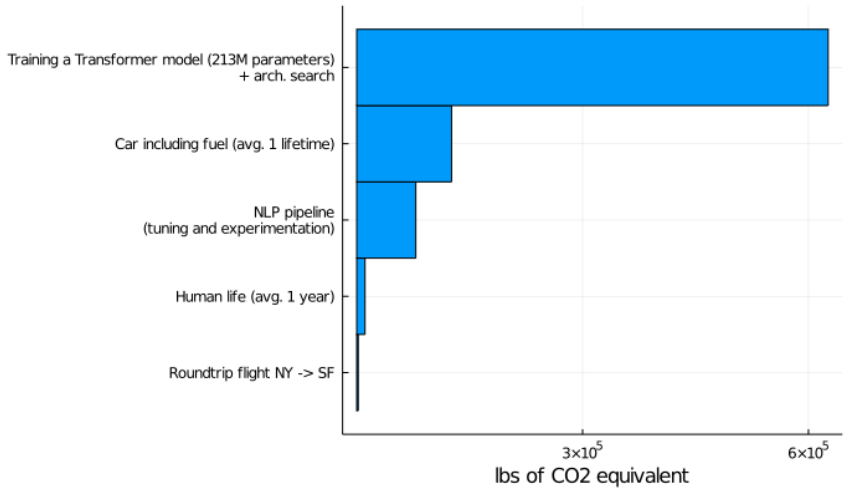


Figure: Consommation d'énergie pour l'entrainement de modèles d'apprentissage profond. Source: Strubell *et al*, 2019 ▸ Link

Table des matières

Table des matières

- modèles basés sur des heuristiques: $\sim$ 1990 – 2005
- modèles bayesiens non-paramétriques : $\sim$ 2005 – 2020
- modèles basés sur des réseaux de neurones: $\sim$ 2015 – 2020

- VAE-HMM: Auto-encodeur avec un Modèle de Markov Caché (MMC) comme distribution *a priori*. ▸ Link
- réseaux de neurones supervisé par des images ▸ Link
- VQ-VAE: Auto-encodeur avec une couche de discrétisation ▸ Link

- Modèles de segmentation des mots ▸ Link 1 ▸ Link 2
- Processus de Dirichlet MMC (2012) ▸ Link
- Processus de Dirichlet MMC avec inférence variationelle (2016) ▸ Link

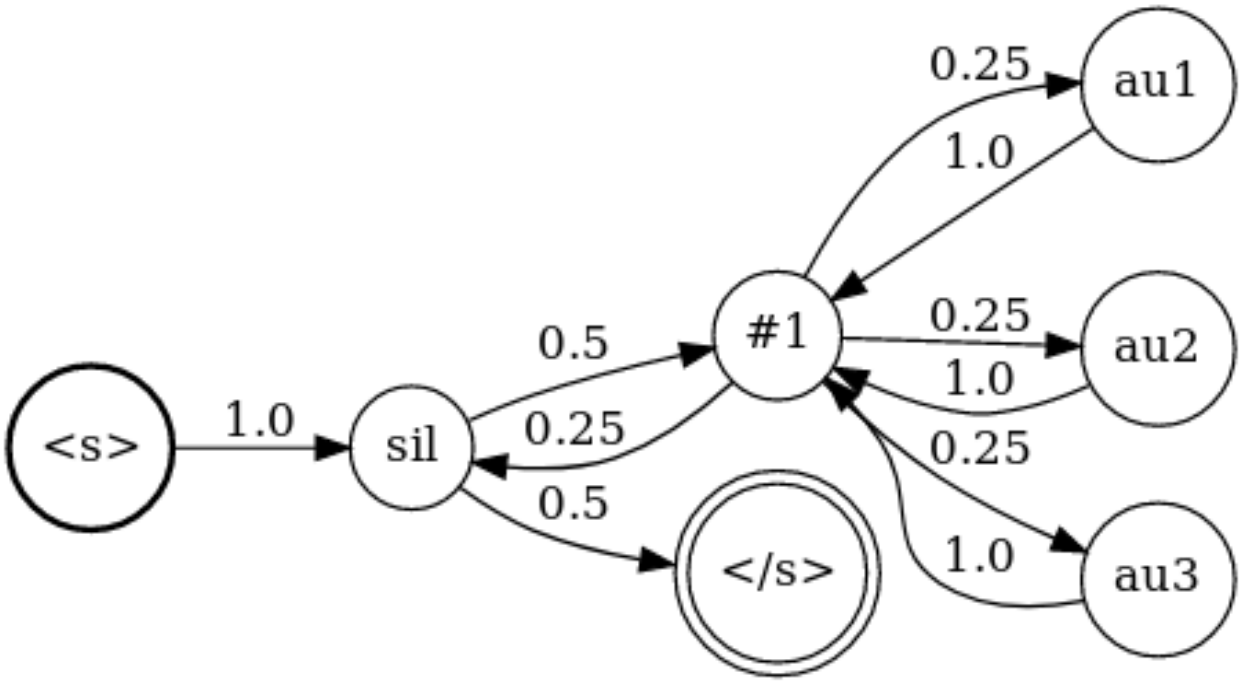
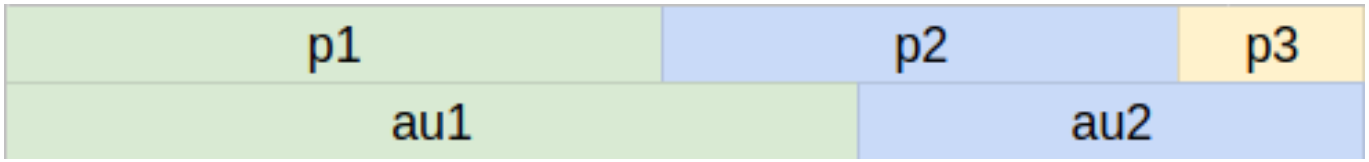

<s>
1.0
sil
0.5
0.25
0.5
#1
0.25
1.0
0.25
1.0
0.25
au1
au2
au3
</s>

Table des matières



- Partitionnement des segments (clustering):
  - Information Mutuelle Normalisé (IMN): $200\frac{MI(X,Y)}{H(X)+H(Y)}\%$
  - Mesure la relation statistique en les UA découvertes et la transcription de comparaison
  - 100 % $\rightarrow$ bijection entre les UAs et les "vrais phones"
  - 0 % $\rightarrow$ les UAs ne donnent aucune information phonétique
- Segmentation:
  - F-score: moyenne harmonique entre la précision et le rappel entre les frontières des segments
  - $\pm$ 20 ms de tolérance

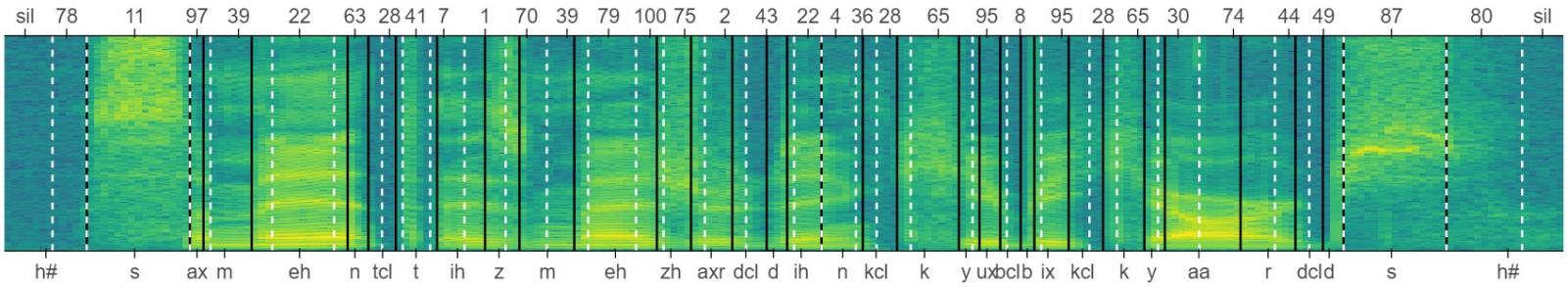
- Data:
  - Mboshi (Congo Brazzaville) - 3-4 heures
  - Yoruba (West Africa - Nigeria) - 10 heures
  - English (TIMIT) - 4 heures
- signal d'entrée: MFCCs + Δ + $\Delta\Delta$

- Version adaptée de "Chorowski et al., 2019" Link
- Processus de Dirichlet MMC Link

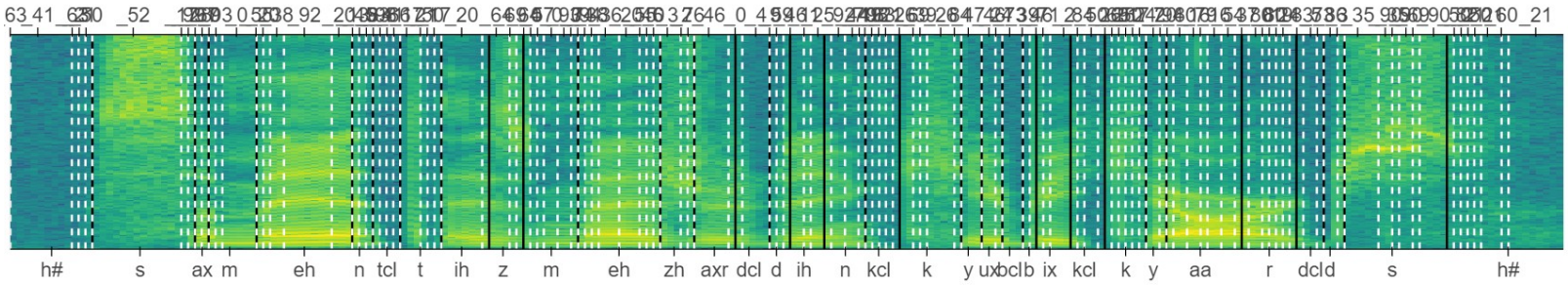
Corpus	Système	IMN	F-Score	# unités
English	VQ-VAE	29.73	38.59	100
English	MMC	**35.47**	**63.03**	95
Mboshi	VQ-VAE 64	26.85	20.22	100
Mboshi	MMC	**36.47**	**47.93**	94
Yoruba	VQ-VAE 64	29.36	7.74	100
Yoruba	MMC	**36.71**	**28.47**	95

Table: Comparaison entre le modèle de base MMC et le VQ-VAE

Exemple



HMM

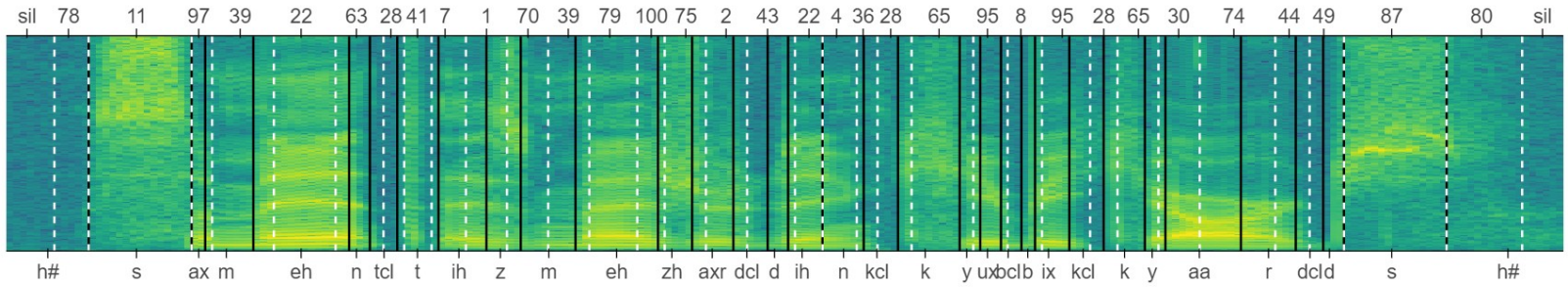


VQ-VAE

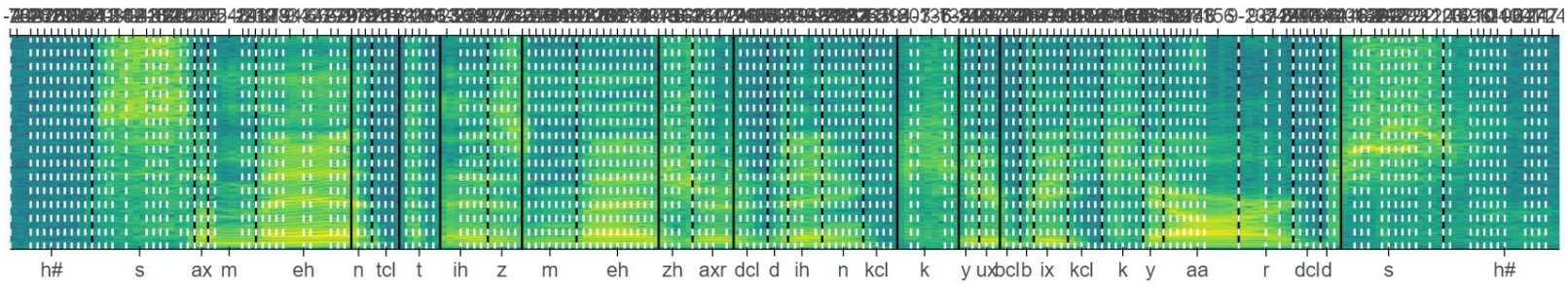
- VQ-WAV2VEC "A. Baevski et al., 2020" trained on 960h of LibriSpeech (unsupervised) ▸ Link
- Processus de Dirichlet MMC ▸ Link

Corpus	Système	IMN	F-Score	# unités
English	VQ-WAV2VEC (Gumbel)	35.20	26.84	12008
English	VQ-WAV2VEC (K-mean)	34.06	25.64	20057
English	MMC	**35.47**	**63.03**	95

Exemple



HMM

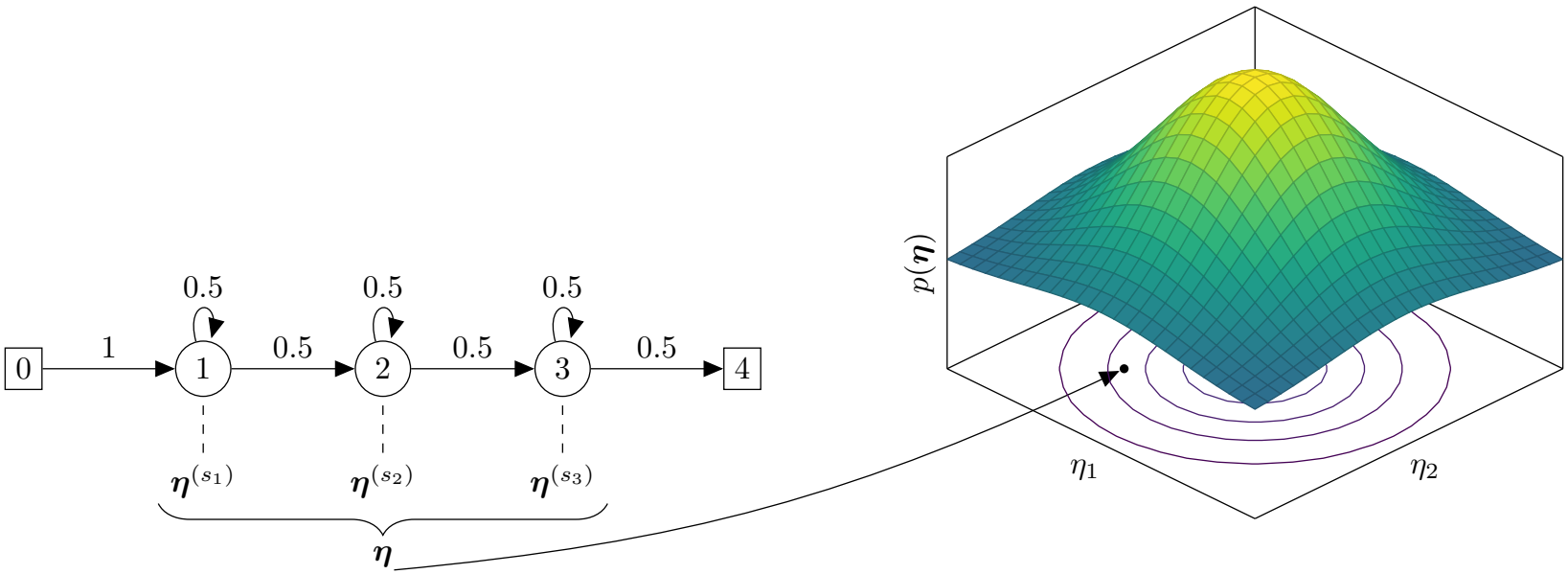


VQ-WAV2VEC

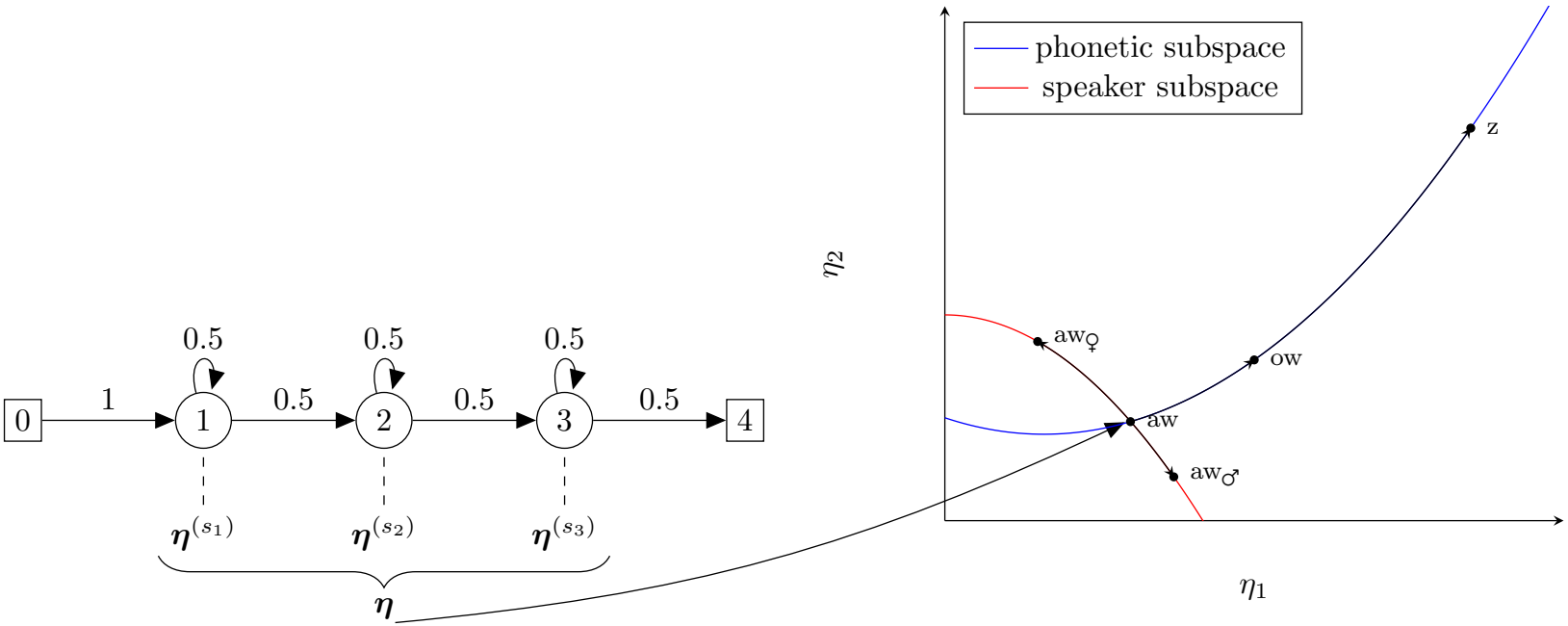
- Les nourissons n'apprennent pas "à partir de rien" (Kuhl *et al*, 1992 Link):
  - Ils ont une sensibilité innée aux langues humaines
  - Avec le temps, ils deviennent spécialisés dans leur langue maternelle
- Choix de design:
  - Le système de DUA doit être "sensible" à la parole à $t = 0$, **il doit posséder de l'information *a priori***
  - Le système de DUA doit s'adapter à la langue cible, **il doit *remettre en question* de l'information fourni en amont**
- Approche proposée: utilisation des modèles de sous-espaces pour implémenter ces propriétés:
  - Le Sous-espace de Modèle de Markov Caché (SMMC) / Subspace Hidden Markov Model (SHMM)
  - Le sous-espace hiérarchiques de Modèle de Markov Caché (SHMMC) / Hierarchical Subspace Hidden Markov Model (H-SHMM)

Table des matières

$$p(\boldsymbol{\eta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\eta})p(\boldsymbol{\eta})}{p(\mathbf{X})} \tag{1}$$



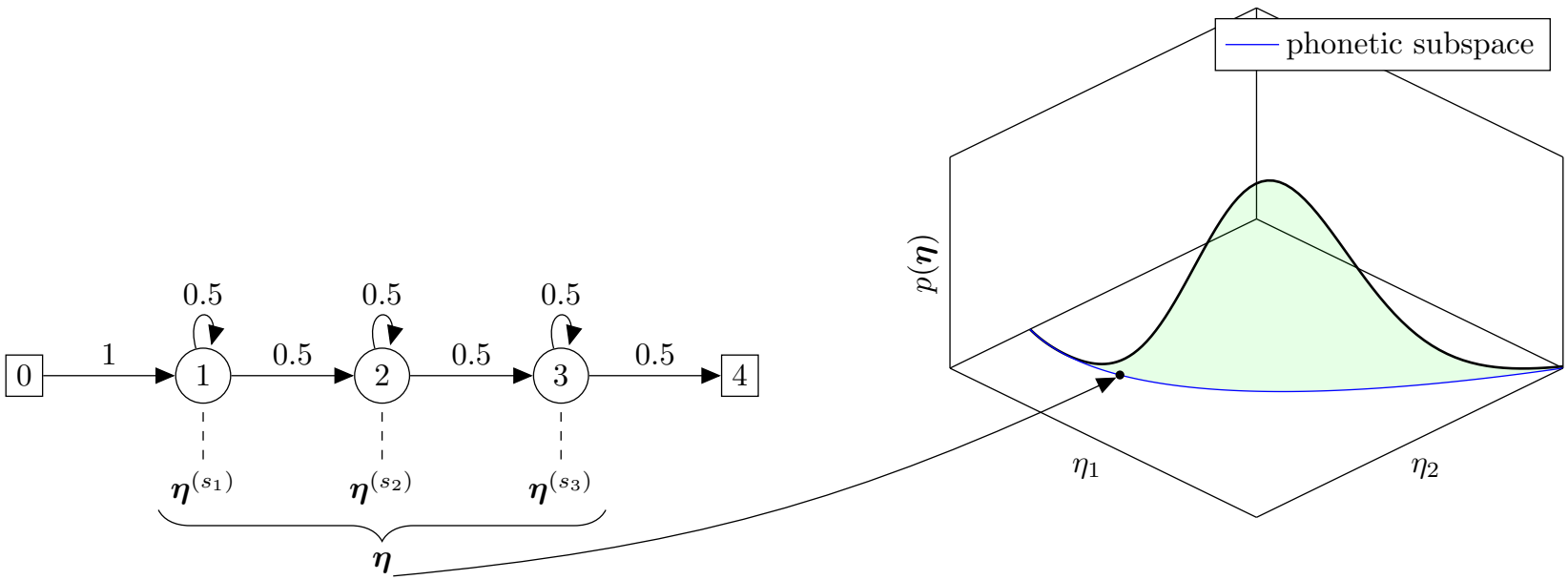
Le sous-espace phonétique



- Nous voulons construire une distribution *a priori* informative sur les paramètres des AU's: $p(\boldsymbol{\eta})$
-

$$\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$

$$\boldsymbol{\eta} = f(\mathbf{W}\mathbf{h} + \mathbf{b}) \tag{3}$$


0.5
0.5
0.5
1
0.5
0.5
0.5
0
1
2
3
4
$\boldsymbol{\eta}^{(s_1)}$
$\boldsymbol{\eta}^{(s_2)}$
$\boldsymbol{\eta}^{(s_3)}$
$\boldsymbol{\eta}$
phonetic subspace
$p(\boldsymbol{\eta})$
η_1
η_2

- On estime les paramètres du sous-espace phonétique $\mathbf{W}, \mathbf{b}$ sur des corpora transcrits
- Le modèle est entraîné par optimisation de la borne inférieure de l'évidence (Evidence Lower-BOund ELBO):

$$\begin{aligned}\mathcal{L} = &\langle \ln p(\mathbf{X}, |\mathbf{z}, \mathbf{W}, \mathbf{b}, \mathbf{h}_{1:T})\rangle_q \\ &- \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z})) \\ &- \mathrm{D}_{\mathrm{KL}}(q(\mathbf{W})q(\mathbf{b}))||p(\mathbf{W})p(\mathbf{b})) \\ &- \mathrm{D}_{\mathrm{KL}}(q(\mathbf{h}_{1:T})||p(\mathbf{h}_{1:T}))\end{aligned}$$

- L'apprentissage est similaire à un algorithme d'*espérance-maximisation*:
  - E-step: algorithme de Baum-Welch pour estimer l'occupations des états de la chaine de Markov
  - M-step: Pas de solution analytique, utilisation de montée de gradient stochastique.

- Les paramètres du sous-espace phonétique $\mathbf{W}, \mathbf{b}$ sont fixés, on apprend simplement le plongement $\mathbf{h}$ sur la langue cible
- Le modèle est entraîné par optimisation de la borne inférieure de l'évidence (Evidence Lower-BOund ELBO):

$$\begin{aligned}\mathcal{L} = &\langle \ln p(\mathbf{X}, |\mathbf{z}, \mathbf{W}, \mathbf{b}, \mathbf{h}_{1:T})\rangle_q \\ &- \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z})) \\ &- \mathrm{D}_{\mathrm{KL}}(q(\mathbf{h}_{1:T})||p(\mathbf{h}_{1:T}))\end{aligned}$$

- L'apprentissage est similaire à un algorithme d'*espérance-maximisation*:
  - E-step: algorithme de Baum-Welch pour estimer l'occupations des états de la chaine de Markov
  - M-step: Pas de solution analytique, utilisation de montée de gradient stochastique.

Link

- Data:
  - Langues sources (transcrites)
    - Français, Allemand, Polonais, Espagnol (Globalphone)
    - 3-4 heures par échantillon pour chaque lange
  - Langues cibles (non transcrites)
    - Mboshi (Congo Brazzaville) - 3-4 heures
    - Yoruba (West Africa - Nigeria) - 10 heures
    - English (TIMIT) - 4 heures
- signal d'entrée: MFCCs + Δ + $\Delta\Delta$:

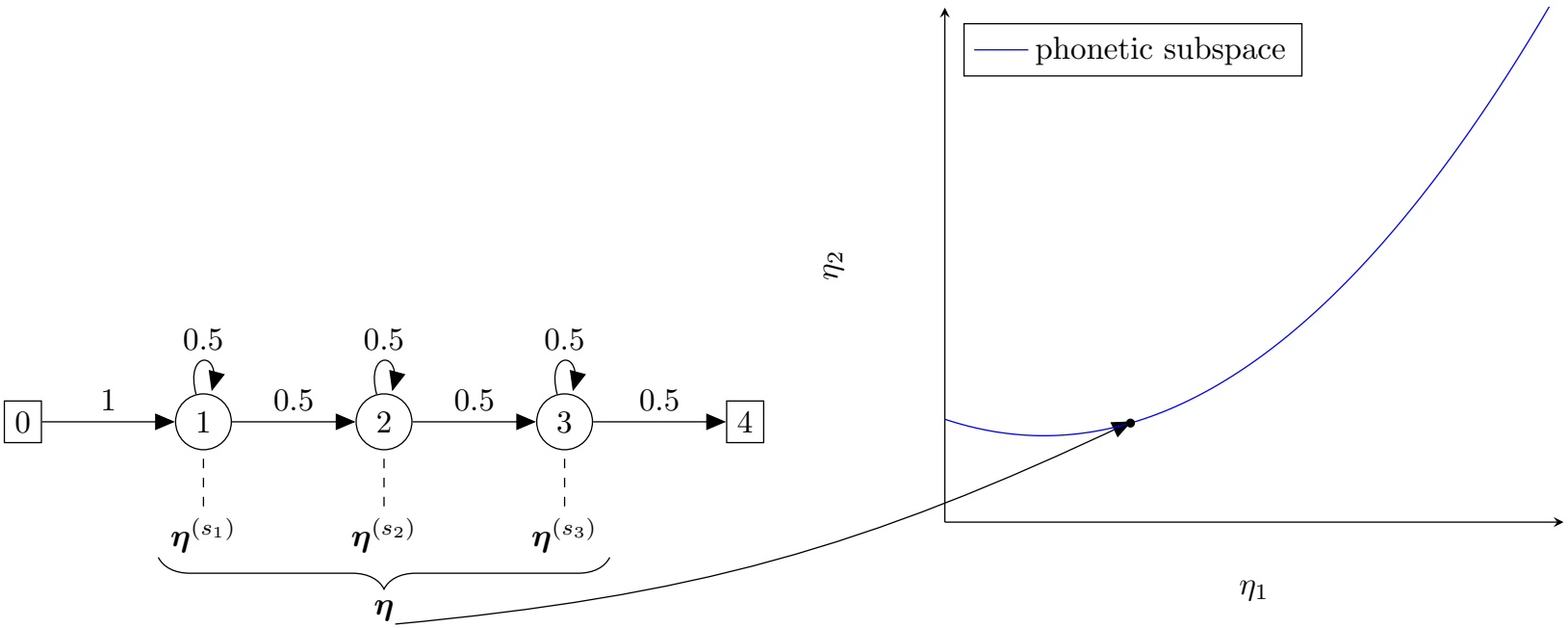
Corpus	Système	Entraîné	IMN	F-Score
English	MMC	non	1.74	0.20
English	SMMC	non	20.83	58.94
Mboshi	MMC	non	1.65	0.02
Mboshi	SMMC	non	21.0	39.28
Yoruba	MMC	non	4.43	1.26
Yoruba	SMMC	non	26.1	27.45

Corpus	Système	Entraîné	IMN	F-Score
English	MMC	non	1.74	0.20
English	MMC	oui	35.47	63.03
English	SMMC	non	20.83	58.94
English	SMMC	oui	**39.66**	**75.92**
Mboshi	MMC	non	1.65	0.02
Mboshi	MMC	oui	36.47	47.93
Mboshi	SMMC	non	21.0	39.28
Mboshi	SMMC	oui	**38.42**	**57.26**
Yoruba	HMM	non	4.43	1.26
Yoruba	HMM	oui	35.27	28.83
Yoruba	SMMC	non	26.1	27.45
Yoruba	SMMC	oui	**37.56**	**36.64**

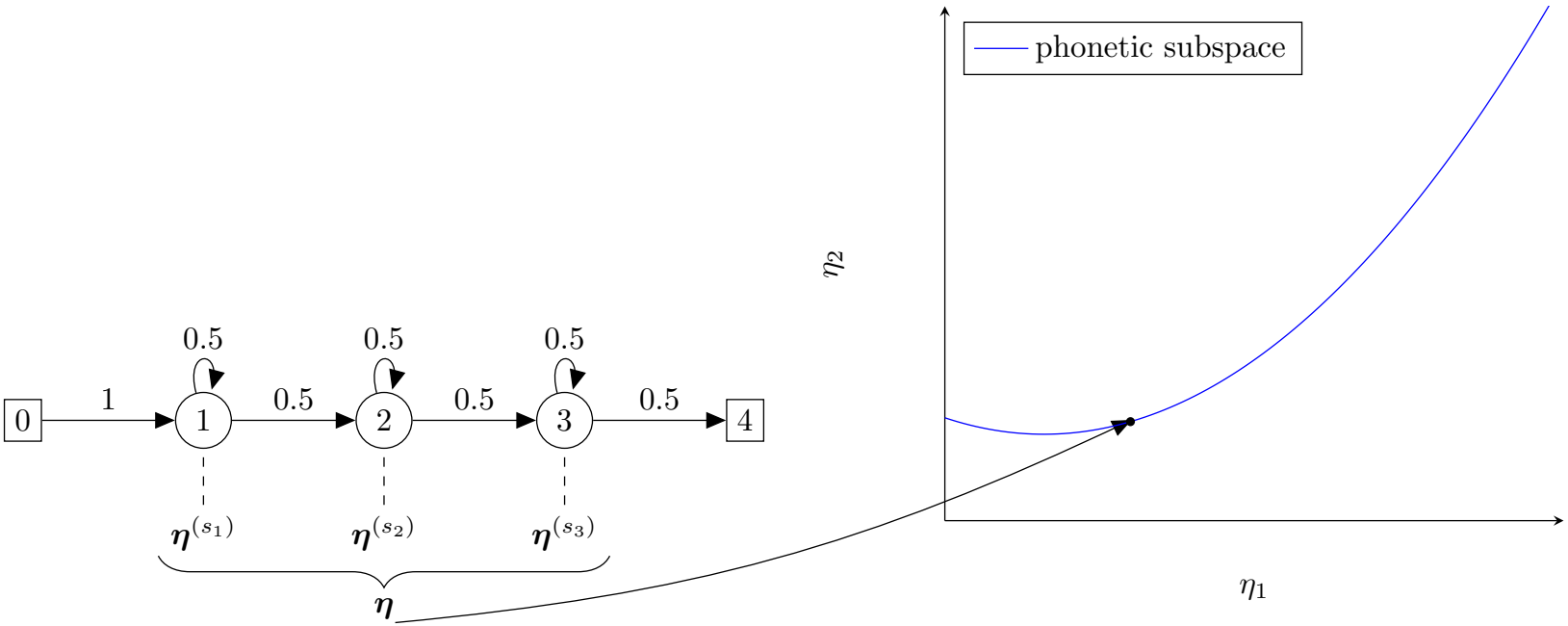
Corpus	Système	Entraîné	IMN	F-Score
English	MMC	non	1.74	0.20
English	MMC	oui	35.47	63.03
English	SMMC	non	20.83	58.94
English	SMMC	oui	**39.66**	**75.92**
English	SMMC (2)	oui	37.46	72.19
Mboshi	MMC	non	1.65	0.02
Mboshi	MMC	oui	36.47	47.93
Mboshi	SMMC	non	21.0	39.28
Mboshi	SMMC	oui	**38.42**	**57.26**
Mboshi	SMMC (2)	oui	35.50	51.28
Yoruba	MMC	non	4.43	1.26
Yoruba	MMC	oui	35.27	28.83
Yoruba	SMMC	non	26.1	27.45
Yoruba	SMMC	oui	**37.56**	**36.64**
Yoruba	SMMC (2)	oui	35.72	31.34

SMMC (2): le sous-espace est ré-entraîné sur la langue cible.

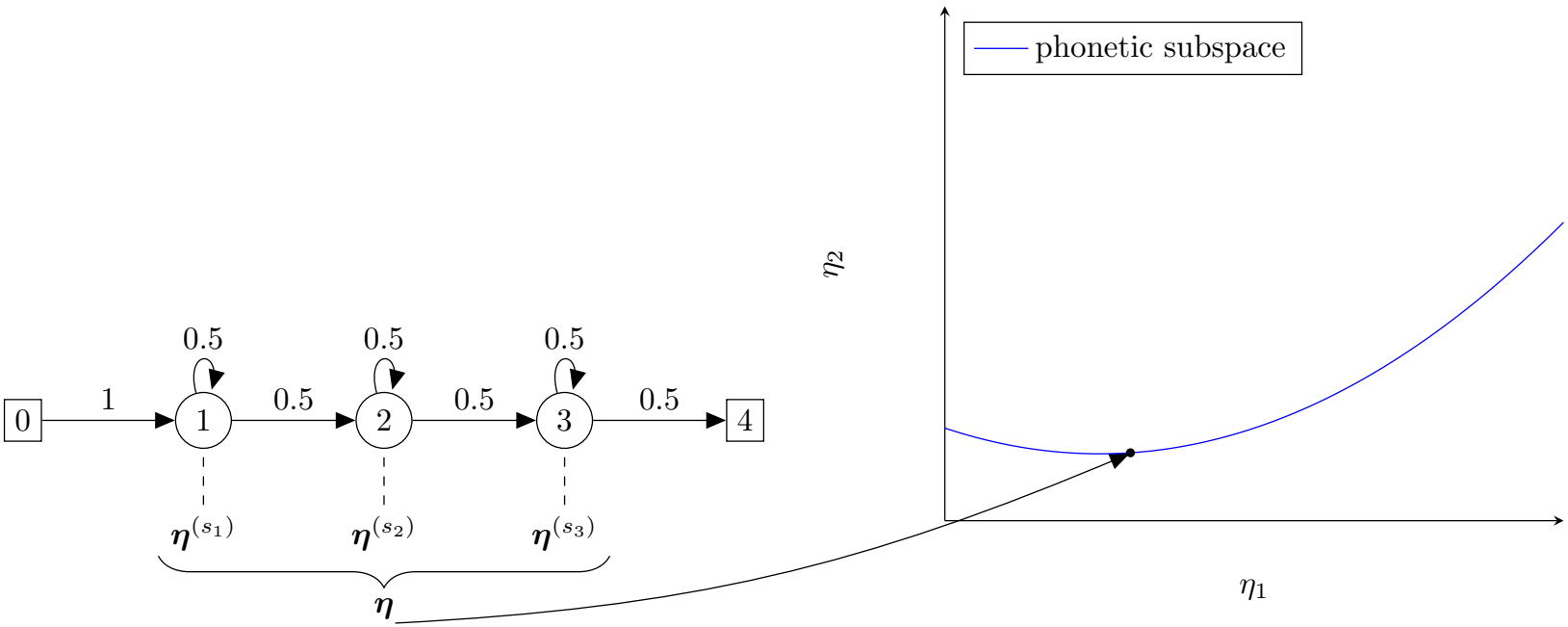
- Nous avons assumé auparavant que le sous-espace phonétique est connu et fixe durant la DUA
- Le sous-espace est le même pour toutes les langues cibles



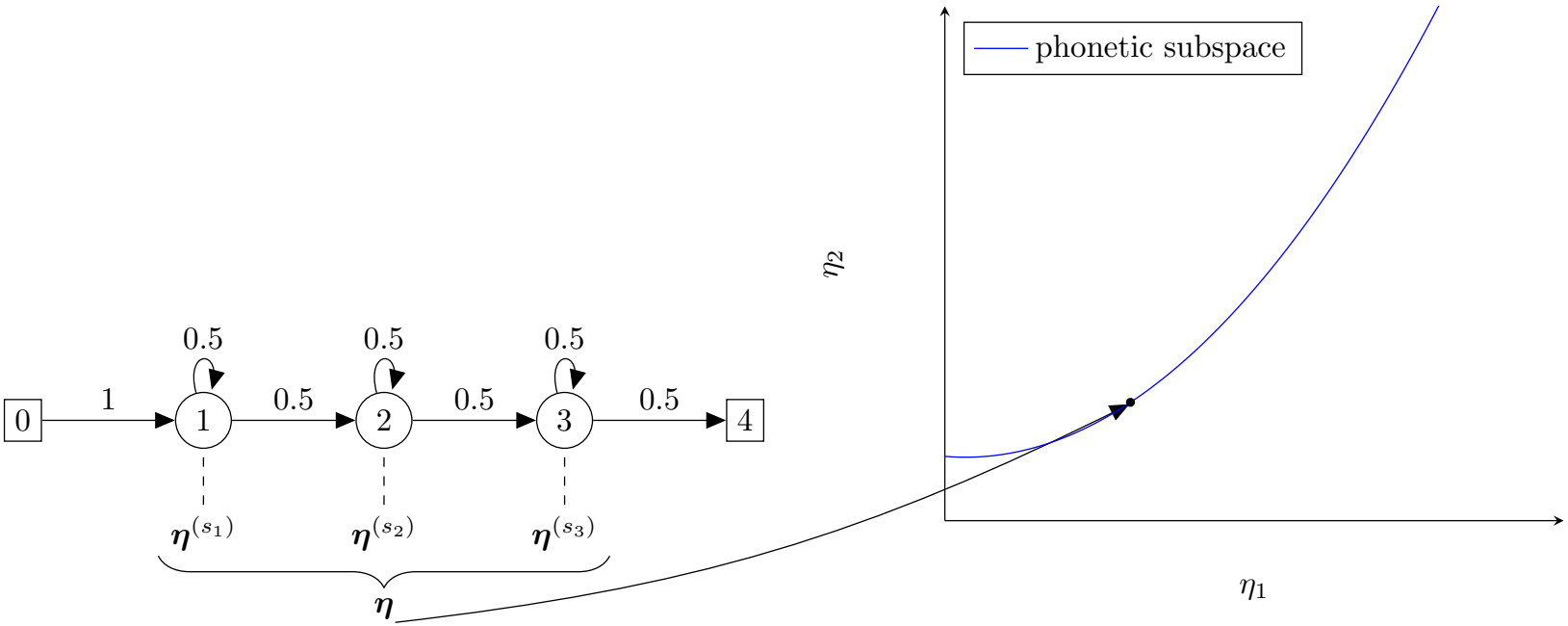
- Nous voulons adapter ce sous-espace pour chaque langue séparément



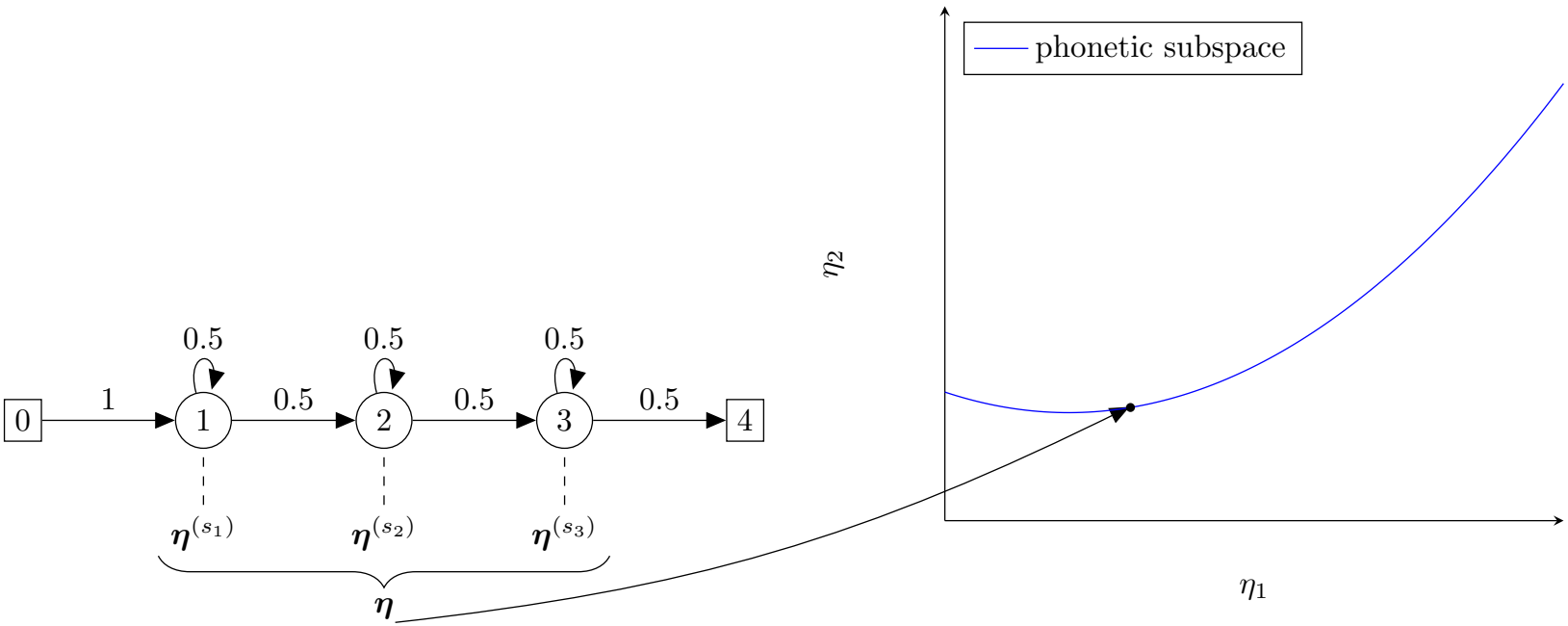
- Nous voulons adapter ce sous-espace pour chaque langue séparément



- Nous voulons adapter ce sous-espace pour chaque langue séparément



- Nous voulons adapter ce sous-espace pour chaque langue séparément



- On construit une distribution *a priori* informative des sous-espaces possibles: $p(\mathbf{W}, \mathbf{b})$
-

$$\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{4}$$

$$\mathbf{W} = \sum_{i=1}^{Q} \alpha_i \mathbf{M}_i \tag{5}$$

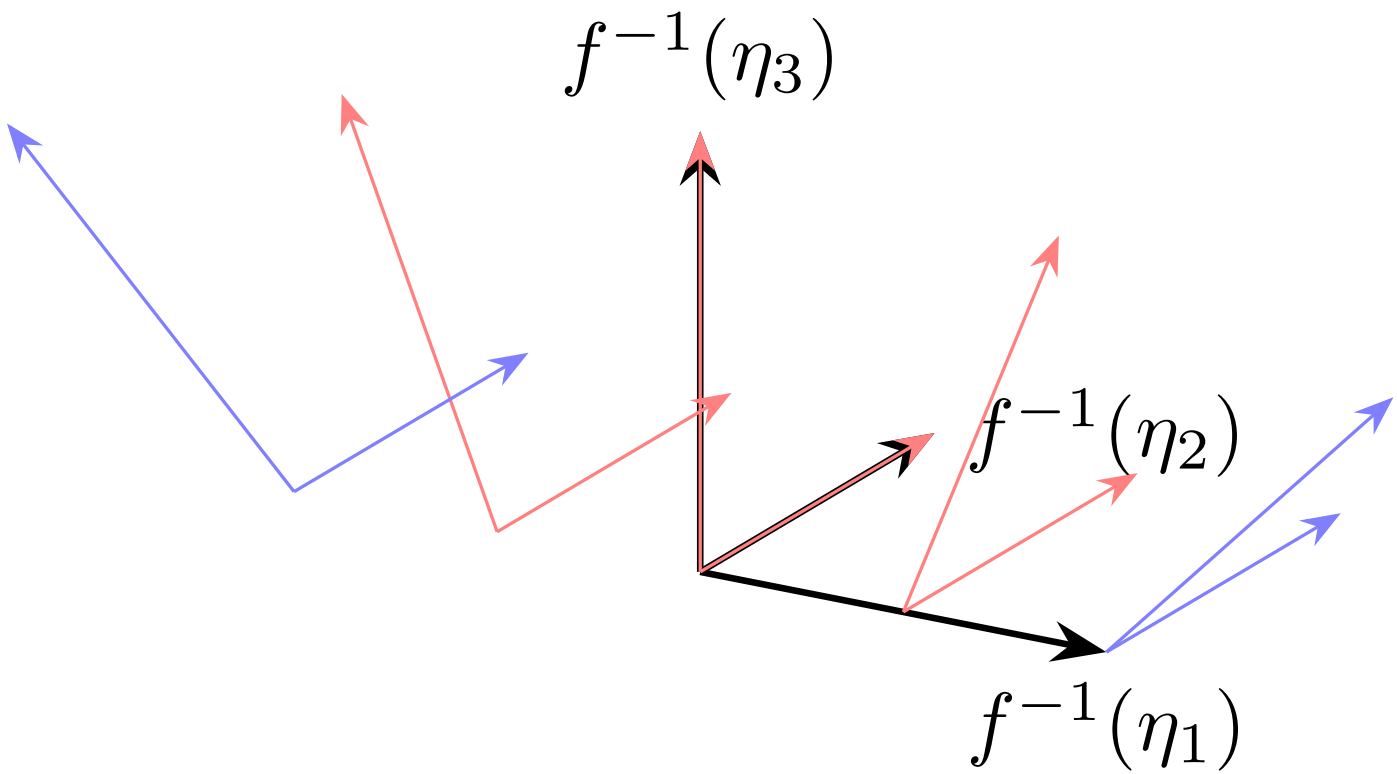
$$\mathbf{b} = \sum_{i=1}^{Q} \alpha_i \mathbf{m}_i \tag{6}$$

$$\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{7}$$

$$\boldsymbol{\eta} = f(\mathbf{W}\mathbf{h} + \mathbf{b}) \tag{8}$$

$$\boldsymbol{\alpha} = \left[\alpha_1, \alpha_2\right]^\top \tag{9}$$

$$\mathbf{W} = \alpha_1 \mathbf{M}_1 + \alpha_2 \mathbf{M}_2 \tag{10}$$



- On estime les paramètres du "méta-sous-espace" $\mathbf{M}, \mathbf{m}, \boldsymbol{\alpha}$ sur les corpora transcrits
- Le modèle est entraîné par optimisation de la borne inférieure de l'évidence (Evidence Lower-BOund ELBO):

$$\begin{aligned}\mathcal{L} = \langle \ln p(\mathbf{X}, |\mathbf{z}, \mathbf{M}_{1:Q}, \mathbf{m}_{1:Q}, \mathbf{h}_{1:T}, \boldsymbol{\alpha}) \rangle_q \\ - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z})) \\ - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{M}_{1:Q})q(\mathbf{m}_{1:Q})||p(\mathbf{M}_{1:Q})p(\mathbf{m}_{1:Q})) \\ - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{h}_{1:T})||p(\mathbf{h}_{1:T})) \\ - \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\alpha})||p(\boldsymbol{\alpha}))\end{aligned}$$

- L'apprentissage est similaire à un algorithme d'*espérance-maximisation*:
  - E-step: algorithme de Baum-Welch pour estimer l'occupations des états de la chaine de Markov
  - M-step: Pas de solution analytique, utilisation de montée de gradient stochastique.

- Les paramètres du "méta-sous-espace" $\mathbf{M}, \mathbf{m}$ sont fixés, on apprends les plongements phonétiques $\mathbf{h}$ et le plongement de la langue $\boldsymbol{\alpha}$
- Le modèle est entraîné par optimisation de la borne inférieure de l'évidence (Evidence Lower-BOund ELBO):

$$\begin{aligned}\mathcal{L} = \langle \ln p(\mathbf{X}, |\mathbf{z}, \mathbf{M}_{1:Q}, \mathbf{m}_{1:Q}, \mathbf{h}_{1:T}, \boldsymbol{\alpha}) \rangle_q \\ - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z})) \\ - \mathrm{D}_{\mathrm{KL}}(q(\mathbf{h}_{1:T})||p(\mathbf{h}_{1:T})) \\ - \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\alpha})||p(\boldsymbol{\alpha}))\end{aligned}$$

- L'apprentissage est similaire à un algorithme d'*espérance-maximisation*:
  - E-step: algorithme de Baum-Welch pour estimer l'occupations des états de la chaine de Markov
  - M-step: Pas de solution analytique, utilisation de montée de gradient stochastique.

Link

- Data:
  - Langues sources (transcrites)
    - Français, Allemand, Polonais, Espagnol (Globalphone)
    - 3-4 heures par échantillon pour chaque langue
  - Langues cibles (non transcrites)
    - Mboshi (Congo Brazzaville) - 3-4 heures
    - Yoruba (West Africa - Nigeria) - 10 heures
    - English (TIMIT) - 4 heures
- Features:
  - traditional features: MFCCs + Δ + $\Delta\Delta$

Corpus	Système	IMN	F-Score
English	MMC	35.47	63.03
English	SMMC	39.66	75.92
Mboshi	MMC	36.47	47.93
Mboshi	SMMC	38.42	57.26
Yoruba	MMC	35.27	28.83
Yoruba	SMMC	37.56	36.64

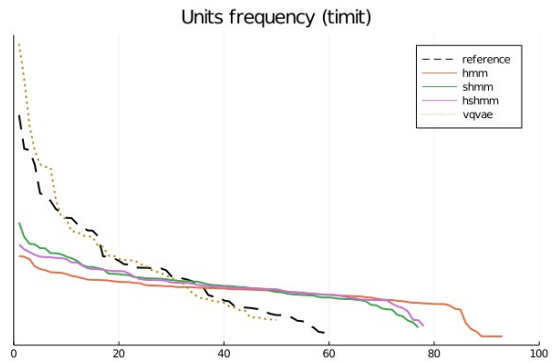
Corpus	Système	IMN	F-Score
English	MMC	35.47	63.03
English	SMMC	39.66	75.92
English	SMMC (2)	37.46	72.19
Mboshi	MMC	36.47	47.93
Mboshi	SMMC	38.42	57.26
Mboshi	SMMC (2)	35.50	51.28
Yoruba	MMC	35.27	28.83
Yoruba	SMMC	37.56	36.64
Yoruba	SMMC (2)	35.72	31.34

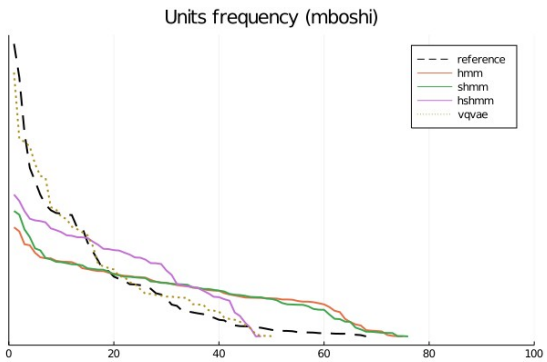
SMMC (2): le sous-espace est ré-entraîné sur la langue cible.

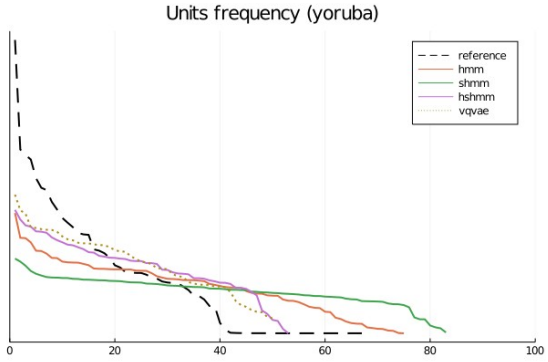
Corpus	Système	IMN	F-Score
English	MMC	35.47	63.03
English	SMMC	39.66	75.92
English	SMMC (2)	37.46	72.19
English	SHMMC	**40.56**	**78.32**
Mboshi	MMC	36.47	47.93
Mboshi	SMMC	38.42	57.26
Mboshi	SMMC (2)	35.50	51.28
Mboshi	SHMMC	**41.17**	**60.82**
Yoruba	MMC	35.27	28.83
Yoruba	SMMC	37.56	36.64
Yoruba	SMMC (2)	35.72	31.34
Yoruba	SHMMC	**37.88**	**38.44**

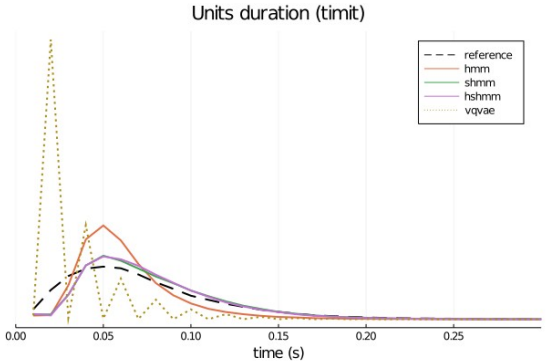
SMMC (2): le sous-espace est ré-entraîné sur la langue cible.

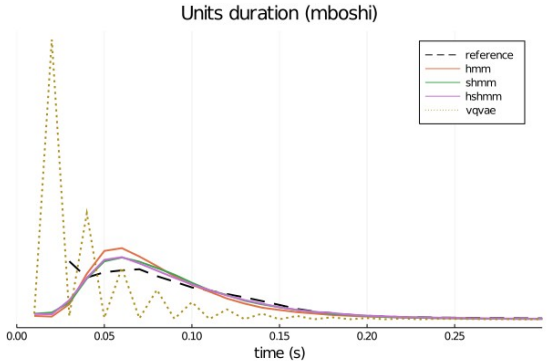
Quelques statistiques

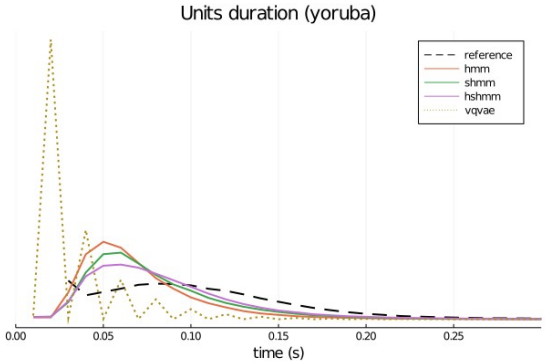




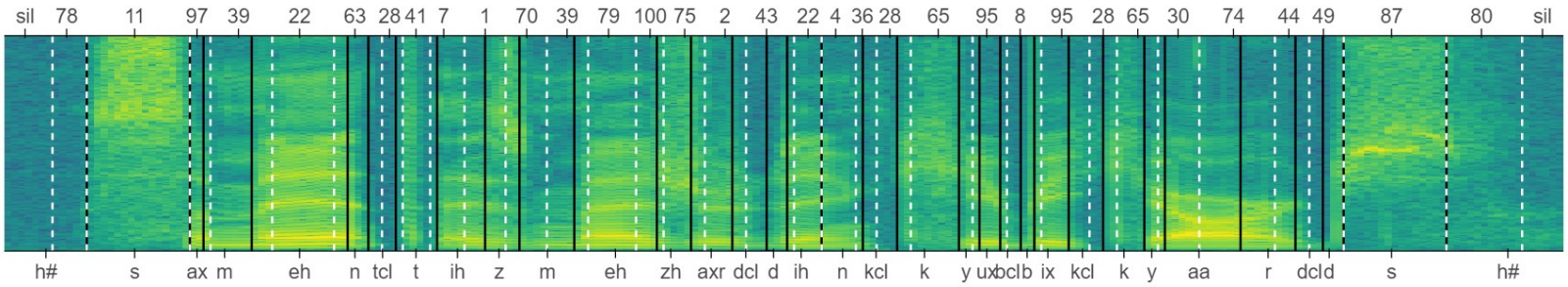




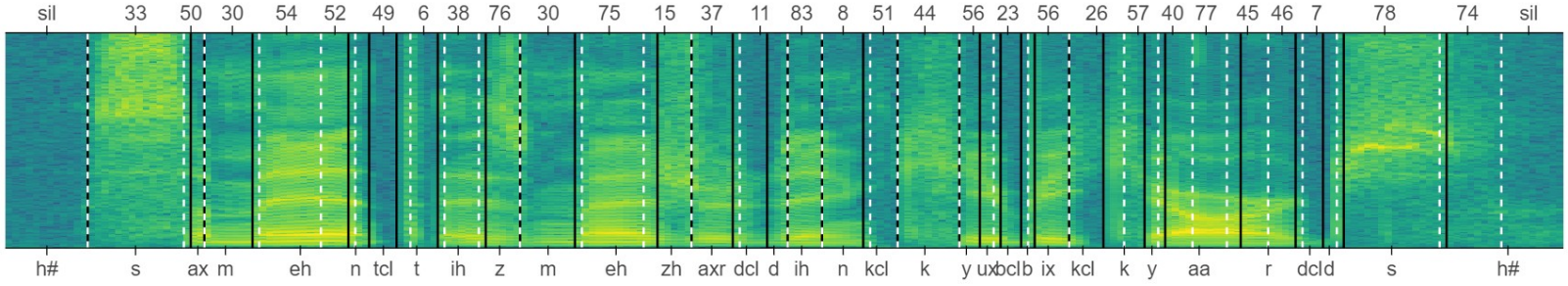




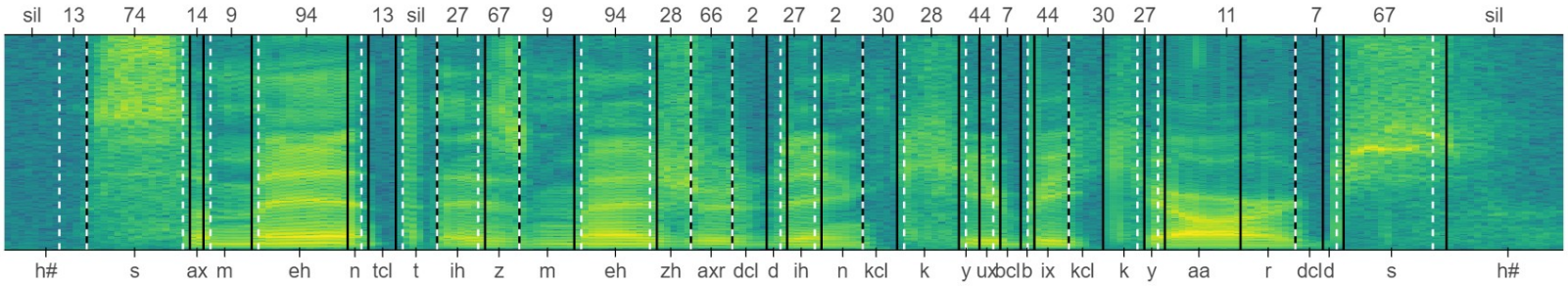
Exemple



HMM



SHMM



HSHMM

Table des matières

En résumé

- Nous avons proposé deux nouveaux modèles pour la DUA:
  - Sous-espace MMC / Subspace Hidden Markov Model
  - Sous-espace hiérarchique MMC / Hierarchical Subspace Hidden Markov Model
- Ces modèles sont inspiré librement de l'apprentissage infantile
- Ils montrent une hausse significative en terme de partitionement et de segmentation
- Le concepte de sous-espace (hierarchique) peut être étendu à une large gamme de modèles
- Reproduire les expériences:
  `https://github.com/beer-asr`

A l'horizon...

- La DUA n'est pas un problème résolu !
- Nos modèles souffrent de la forte variabilité de la parole
- Principaux axes de recherche futurs
  - Modèle acoustique: explorer d'autres modèle génératifs que MMC
  - Modèle de langue: la découverte de mots
- En direction du premier modèle qui apprends la parole comme les humains...

Références I

- Yusuf, Bolaji, Lucas Ondel, Lukás Burget, Jan Cernocký, and Murat Saraclar (2020). "A Hierarchical Subspace Model for Language-Attuned Acoustic Unit Discovery". In: CoRR abs/2011.03115. arXiv: 2011.03115. URL: https://arxiv.org/abs/2011.03115.
- Lucas Ondel, Hari Krishna Vydana, Lukáš Burget, and Jan Černocký (2019). "Bayesian Subspace Hidden Markov Model for Acoustic Unit Discovery". In: Proc. Interspeech 2019, pp. 261–265. URL: http://dx.doi.org/10.21437/Interspeech.2019-2224.
- Lucas Ondel, Pierre Godard, Laurent Besacier, Elin Larsen, Mark Hasegawa-Johnson, Odette Scharenborg, Emmanuel Dupoux, Lukáš Burget, Francois Yvon, and Sanjeev Khudanpur (2018). "Bayesian Models for Unit Discovery on a Very Low Resource Language". In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5939–5943. URL: https://ieeexplore.ieee.org/document/8461545.


Lucas Ondel, Lukaš Burget, Santosh Kesiraju, and Jan Černocký (2017). "Bayesian phonotactic language model for acoustic unit discovery". In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5750–5754. URL: https://www.fit.vut.cz/research/publication/11472/.en.


Lucas Ondel, Lukáš Burget and Jan Černocký (2016). "Variational inference for acoustic unit discovery". In: Procedia Computer Science 81, pp. 80–86. URL: https://www.sciencedirect.com/science/article/pii/S1877050916300473.

Merci pour votre attention.